Archaeal core gene set
Washington University Genome Center

Author: Makedonka Mitreva
Version: 1.01
Effective Date: 12/15/08

1 Abstract

None

2 Introduction

The archaeal core set is used in testing the completeness of the archaeal draft genomes. The core set comprises of conserved single copy genes from 25 genomes. Coverage statistic is calculated by the percentage of the core set that have orthologs in the draft genome.

3 Requirements

3.1 Data requirements

Input fasta file: Input is a protein fasta format file or the assembly sequences.

Core genes: This is the fasta file containing of 2,600 core genes.

Core genes cluster file: Each line of the cluster file contains core genes that are orthologs. There are 104 groups and each group contains 25 genes.

3.2 Software requirements

perl : Any Perl installation

blast or other alignment program : NCBI BLAST or WU- BLAST or other alignment sotware

get_coregroups_coverage.pl : Perl script to calculate coverage of core genes

3.2 Compute requirements

None

4 Procedure

1. **Alignments to core genes**

   Using the core genes file as query and the input fasta file as subject, obtain alignments to the 2,600 core genes. Any alignment software can be used.

   (Example: blastx with parameters " hitdist=40 wordmask=seg postsw topcomboN=1")

2. **Filtering alignments**

   All alignments that have at least than 50% identity over at least 70% of the length of the core genes are considered valid. The core genes that have alignments meeting this criteria are collected to be used as input for the next step.

3. **Coverage of core groups**

   The list of core genes that had homology to the input fasta sequences with 50% identity over 70% length are passed on to this script – get_coregroups_coverage.pl

   Script run with the fo]tllowing parameters –

   get_coregroups_coverage.pl
   -coregroups <cluster file>
   -gene_list <file with core genes with valid alignments>

   The gene list is the list of core genes from step 2.
   The coregroups parameters is the core group cluster file.

5 Implementation

The script takes the orthologs for the core genes in the draft genomes as input and searches against the core groups to identify the number of groups that have at least one of the genes. The percentage of core groups identified is printed as output.

Author: Makedonka Mitreva
Version: 1.01
Effective Date: 12/15/08

6 Discussion

To build the archaeal core set, 112,992 sequences from 52 archaeal genomes (16 Crenarchaeota, 34 Euryarchaeota, 1 Nanoarchaeon and 1 Korarchaeota; downloaded from Genbank on 05/13/2008) were clustered using OrthoMCL[1]. OrthoMCL constructs putative ortholog groups by applying a markov cluster algorithm on BLAST results obtained by searching all sequences against themselves. A total of 11,410 orthologous groups were identified from these sequences with default parameter settings (IF=1.5). Out of these groups, 119 groups with 6,445 genes comprised of genes from all of the 52 species and 84 of these groups were single-copy orthologs, i.e. only one gene from each species. This dataset was compared to a previous core set reported by Makarova et al. [2] which constructed 166 Archaeal Clusters of Orthologous Genes (arCOGs) from 41 genomes (13 Crenarchaeota, 27 Euryarchaeota and 1 Nanoarchaeon). Thirty-nine of these 41 genomes were included in our study. Of these 166 arCOGs, 86 are single-copy groups from all the 41 genomes. The 84 orthologous gene groups identified by OrthoMCL were found to be a subset of these 86 arCOGs.

A similarity dendrogram based on the 84 WashU groups was constructed and these sequences were aligned using MUSCLE [3] and the alignments were concatenated to form a combined alignment. The Protdist software from Phylip[4] was used to calculate the distance matrix based on this combined alignment and then the FITCH program, also from Phylip, was used to calculate the phylogeny.

These 52 species include multiple representatives from closely related strains or species. In order to get a reduced representative gene set, some of the closely related species were removed from the orthologous gene set. Distances calculated from the Protdist program were used for this procedure and a distance cut-off was applied to reduce the number of genomes. Among species pairs that had distances lesser than the cut-off, one of the pair was randomly discarded.Three cut-off parameters were tested - 0.1(46 genomes remained), 0.25 (37 genomes) and 0.5 (25 genomes). The cut-off of 0.5 was chosen based on the similarity tree constructed from the remaining genomes as this included all the major

branches in the dendrogram and removed branches that were small and closely related. All the 25 species remaining with this cut-off had distances greater than or equal to 0.5.

Orthologs from these 25 genomes were constructed from OrthoMCL and 5,241 groups were identified. From these groups, 133 had a representative from all 25 species and 104 of them were single-copy groups. These 2,600 genes from the single copy 104 groups represent the core archaeal set (separate file, list the gis of the core genes).

Coverage of the draft genomes can be estimated by identifying which of the 104 groups have orthologs. Evaluating if genes from these core groups are present in the draft genomes gives a good indication of the completeness and coverage of the draft genomes.

A more detailed description of Archaeal core gene selection and evaluation can be found in Abubucker & Mitreva, Identification of Archaeal Core set.

7 Related Documents & References

[1] Li Li, Christian J Stoeckert, Jr. and Davis S. Roos (2003), OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes, *Genome Res. 2003. 13: 2178-2189*

[2] Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea, Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV, *Bioloogy Direct 2007 Nov 27;2:33*

[3] MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, 32(5), 1792-97

[4] PHYLIP (Phylogeny Inference Package) version 3.6, Felsenstein, J. 2005. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

8 Revision History

This is an HMP_specific requirement, not included in the SIGS submission. Please be sure to update this when any changes as made, to help the DACC organize SOPs.

Version	Author/Reviewer	Date	Change Made
1.01	Makedonka Mitreva	12/15/08	Establish SOP